融合句法特征和句法相似度的网络舆情突发事件识别方法研究

■ 陈健瑶 翟姗姗 夏立新 刘德印
华中师范大学信息管理学院 武汉 430079

摘 要：[目的/意义] 快速、准确地从突发网络舆情文本中识别事件。[方法/过程] 提出一种融合句法特征和句法相似度的网络舆情突发事件识别方法。结合句法特征提出面向事件的句法特征提取方法，利用事件语义标注和句法特征提取方法构造事件句法特征库，通过计算待测文本与句法库的句法相似度来识别网络舆情突发事件。[结果/结论] 以新型冠状病毒肺炎疫情为例，所提出网络舆情突发事件识别方法在该舆情下的最优相似度为 0.93，在此相似度下从一段新的文本中识别出 160 个事件和 30 个非事件，F1 值达到了 0.848。通过方法测评证明网络舆情突发事件识别方法在利用句法相似度识别事件和进行相同相邻词性合并等方面创新的有效性。

关键词：网络舆情 事件识别 句法特征 句法相似度
分类号：G250.2
DOI：10.13266/j.issn.0252-3116.2021.09.005

1 引言

中国互联网络信息中心(CNNIC)发布的第 44 次《中国互联网络发展状况统计报告》显示[1]，截至 2019 年 6 月，我国网民规模达 8.54 亿，互联网普及率达 61.2%，较 2018 年底提升 1.6 个百分点。网络的普及与平民化使得公众对舆情事件的关注和回应更为便捷，所产生的网络舆情内容更为丰富。在此背景下，快速、准确地从网络舆情突发事件文本中识别能反映公众态度、舆论走向的事件，并为政府实施引导策略提供针对性的参考意见，成为网络舆情研究领域的一项重大挑战。

网络舆情突发事件是指能反映公众对网络突发社会问题不同看法的事件。事件具备抽象性、广义性和语义完备性等特点，其表示形式为事件三元组 $E=(S,P,O)$，其中 P 是触发词，S 是施事者，O 是受事者[2]。一个完整的事件必须要包含触发词，触发词决定了事件的类型，施事者和受事者可以部分忽略，例如“台风登陆(S,P)”“看电影(P,O)”“贵州凉山爆发山火(S,P,O)”都可以称之为事件。

网络舆情突发事件识别任务主要研究从非结构化的社交媒体数据中识别包含事件元素的结构化事件文本。从 2005 年起，事件抽取被纳入 ACE 评测会议[3]，事件识别是事件抽取任务的重要组成部分。事件抽取可分为主题事件抽取和元事件抽取。主题事件是指与某个主题相关的一组事件，它由一个核心事件和所有与之直接相关的事件或活动组成[4]；元事件主要描述了参与动作事件的主要成分结构，其通常使用动名词表示动作的发生或状态的改变。中文事件识别技术在国内外学者的研究下也取得了长足的进步[5]。相较于英文清晰的语句结构，中文词语之间的排列组合更加的复杂和灵活，且词语存在较多的一词多义现象，事件的含义也需要结合上下文语义进行辨别，这给中文事件识别技术带来了一定的困难，如何降低中文文本维度以及词语之间的语义关联成为中文事件识别的一大考验。

为探究适用于中文环境下的网络舆情突发事件识别方法，笔者提出一种融合句法特征和句法相似度的网络舆情突发事件识别模型，以新型冠状病毒肺炎疫情为例，构造网络舆情事件句法特征库，利用语句之间

作者简介：陈健瑶(ORCID:0000-0003-1890-7404)，硕士研究生，E-mail:1041403539@qq.com；翟姗姗(ORCID:0000-0002-2787-0183)，副教授，博士；夏立新(ORCID:0000-0002-4162-2282)，教授，博士，博士生导师；刘德印(ORCID:0000-0002-1769-3160)，硕士研究生。
收稿日期：2020-12-09 修回日期：2021-02-23 本文起止页码：41-50 本文责任编辑：徐健

的句法相似度去识别网络舆情中新的事件。

2 相关研究

不同领域的研究人员对于事件有着不同的定义。事理图谱研究人员将事件定义为抽象、广义和具备完整语义的事件三元组[2]；语言学领域认为事件是由谓语动词及动作发生时间、情况所构成的术语[6]；自动内容抽取(automatic content extraction，ACE)评测会议认为事件是发生在某个特定的时间点或时间段、某个特定的地域范围内，由一个或者多个角色参与的一个或者多个动作组成的事情或者状态的改变[7]。在上述对于事件的定义中，事理图谱研究人员所给定的事件定义因其结构化的特点较为贴合本研究的网络舆情突发事件，因此采用此定义。

2.1 事件识别方法相关研究

事件识别方法主要有两种：基于模式匹配的方法和基于机器学习的方法。基于模式匹配的方法，即在一些模式的指导下进行事件的识别和抽取。模式主要用于指明构成目标信息的上下文约束环境，集中体现了领域知识和语言知识的融合[8]。其方法可分为规则方向的扩展和关系方向的限制，前者倾向于宏观层面的扩展触发词表规模、完善知识库构建等；后者倾向于微观层面的文本信息单元融合、语义一致性推理、语义约束等。目前有学者利用模式匹配进行战争事件抽取[9]。基于机器学习的方法，即运用统计模型进行事件的识别和抽取。该方法在近几年较为主流。常用的学习方法有条件随机场模型[10]、隐马尔科夫模型[11]、支持向量机模型[12]等。贺瑞芳等[10]将事件抽取看作序列标注任务，构建了基于 CRF 多任务学习的中文事件抽取联合模型，针对仅基于 CRF 的事件抽取联合模型的缺陷进行了扩展；刘忠宝等[13]基于 BERT 模型和 LSTM-CRF 模型对历史事件及其组成元素进行抽取。

综上可以看出，关于事件识别方法的研究已取得了较大的进展，能够从文本中精准的识别出事件，但是这些方法大多数依赖于规模较大、范围较全的训练集，从而需要构建知识库进行事件识别，这类方法应用于网络舆情突发事件领域就会面临网络舆情初期训练集语料不足的情况，因此本研究旨在当前事件识别方法研究的基础上提出一个能够适用于网络舆情领域的突发事件识别方法。

2.2 网络舆情突发事件识别研究现状

尉永清[14]等在研究突发事件网络舆情传播规律的基础上，研究事件特征抽取方法和情感特征的突发性，用于识别突发事件，为预测事件发展提供数据支持；武澎[15]等运用博弈分析得出微博中突发事件信息发布者被关注的概率模型，为网络舆情突发事件信息传递关键节点的确立奠定了基础；刘雅姝等[16]利用 LDA 方法对网络舆情突发事件评论数据进行话题划分并构建事件演化话题图谱，用以动态追踪民意了解网络舆情突发事件发展方向；兰月新[17-18]通过建立衍生舆情监测预警模型和突发事件网络舆情信息传播规律模型，为政府实现网络舆情管理和网络舆情预警研究提供参考；张玉亮[19]把突发事件网络舆情划分为生成期、扩散期、衰退平复期 3 个阶段，为政府及其相关部门有效评估突发事件的现实状况，把握突发事件网络舆情发展态势提供比较有效的理论支持和借鉴；陈思菁等[20]利用用户行为特征、网络全局信息以及影响力衰退机制的关键节点动态识别方法，识别突发事件信息传播在不同阶段中的关键节点及其演化特征；李纲等[21]利用主题模型(LDA)与互信息最大熵模型(MaRxE nt-MI)提取事件摘要关键词，进而生成事件摘要；夏立新等[22]从可视化视角出发，从多个维特构造网络舆情事件特征，形成可视化事件摘要。

可以看出，当前已有学者在网络舆情领域进行事件识别和应用研究，但是大多数学者的研究重点在事件传播和舆情发展等方面，并未提出一种能够适用于网络舆情领域的突发事件识别方法，因此，笔者将在当前网络舆情领域事件识别和应用的研究基础上，提出一种网络舆情通用的突发事件识别方法，为后续研究人员基于事件研究网络舆情提供参考。

2.3 句法分析相关研究

句法分析的研究大体分为基于规则的方法和基于统计的方法，前者以语言学理论为基础，后者以某种方式对语法规则和语言形式进行描述[23]。袁里驰[24]建立了一种基于依存关系的句法分析统计模型，将句法分析模型与分词、词性标注模型相结合，取得了良好的实验效果；郭喜跃[25]等将句法特征、语义特征融入依存句法关系、核心谓词、语义角色标注等特征进行实体关系抽取，实验结果表明了融入句法特征后方法的有效性；徐飞等[26]利用 BiLSTM-CRF 模型对食品事件进行词性标注，取得了较好的试验结果；胡宝顺[27]等提出一种新的基于句法结构特征分析及分类技术的答案提取算法，实验结果证明基于句法结构特征的方法性能优于目前典型的算法；陈永波[28]等提出简单边优先与 SVM 相结合的依存句法分析算法，实验结果证明对于复杂名词短语的依存句法分析，算法准确率比简单

边优先算法有明显提高。

学者们对于句法分析的研究与应用证明了句法在表达文本特征方面的有效性，基于此，笔者认为能表达特定事件的中文文本存在着一定的句法模式，在网络舆情初期训练集语料不足的情况下，句法特征能够代替文本特征进行事件识别。使用事件句法来识别事件，能有效降低中文文本的维度，大大降低了事件识别的工作量和复杂程度，同时，该方法可以降低对特定舆情领域词典的依赖，使事件识别方法得到更广泛的应用。

3 基于句法特征提取的句法相似度度量

基于句法特征提取的句法相似度度量主要分为两个子模块：①面向事件识别的语句句法特征提取，事件句法包含了事件框架下的事件语义逻辑；②基于句法特征的事件句法相似度计算方法。两个事件句法相似度越高，说明两个事件在句法语义结构层面越相似。

3.1 面向事件识别的语句句法特征提取

利用分词工具对事件文本进行分词和词性标注，以事件语句“土耳其再次向一核大国开火”为例，对其进行分词和词性标注后得到文本向量：

E = ["土耳其":n, "再次":d, "向":p, "一":m, "核大国":n, "开火":v]

对文本向量 E 进行句法特征提取后得到句法特征向量：

$$P = [n, d, p, m, n, v]$$

通过语句句法特征提取，事件的表征方式由文本向量 E 转换为句法特征向量 P，这就使得事件识别的维度从文本特征转换到句法特征。但这也存在一个问题，词性的种类远远小于词语的种类，许多不同的词语会具有相同的词性，造成了句法特征向量冗余，同时因不同口语化的表达方式，同一事件可能使用多种语言表达方式，为降低这种冗余和句法特征的复杂性，笔者以“相邻相同词性合并”的方式来降低相同事件的句法种类和向量维度，例如，对于句法：

$$P = [n, n, d, p, m, m, n, v]$$

将其简化为：

$$P = [n, d, p, m, n, v]$$

合并相同词性的目的在于对句法种类进行泛化，即默认相同词性的相邻词语表达的语义特征相同，使模型即使在缺少大量文本训练集的情况下，也能发挥出最大优势。同时，由于存在“一词多义”“一义多词”等现象，部分经分词工具标注后的句法存在语义冲突进而导致误差，为减少这种误差，笔者拟采用人工检查的方式将这些冲突的句法模式识别出来，人工通过事件语义关系识别错误句法，并形成错误句法模式词典，此词典放置于后续网络舆情突发事件识别模型中，当识别出待测句法存在于错误句法模式词典中时，说明该句法为错误句法，直接判定为非事件。人工检查错误句法模式的工作量在初期工作量与事件标注工作量相同，但随着错误句法模式词典规模变大，产生语义冲突的句法会越来越少，相应工作量也会越来越少。相较其他减少误差的方法，人工检查因其便捷可控的特点更为适合句法特征提取。事件句法模式提取的具体过程如算法 1 所示：

算法 1：语句句法特征提取

输入：sentences[0..n－1]：包含 n 条待处理语句(sentence)的数组；f1(sentence)：对文本进行分词的函数；f2(word)：对语词进行词性标注的函数；f3(pattern)：对已提取句法进行相邻相同词性合并；

输出：patterns 事件句法集

```
function Pattern(sentences[0..n－1]: array of sentence; f1: function; f2: function; f3: function): patterns;
var
  words[0..m－1]:包含 m 个词的数组;
  nominal:词性标注序列;
begin
  for i←0 to n－1 do
    pattern ← null
    words[0..m－1] ← f1 (sentences[i])
    for i←0 to m－1 do
      nominal ← f2(words[i])
      pattern ← pattern + nominal
    pattern ←f3(pattern)
    if pattern not in patterns then
      patterns ← patterns + pattern
    end if
  return patterns
end
```

3.2 基于句法特征的事件句法相似度计算方法

文本余弦相似度是一种常用的文本相似度度量标准，传统的文本向量余弦相似度能够表达两个文本文档之间的相似度，通过词向量间距来判断两个文档的亲疏关系。笔者对文本余弦相似度进行一定的修改并将其应用于事件句法相似度计算，句法特征向量相似度能够从语义层面表达两个事件在句法逻辑方面的相似性。将待识别事件句法 $P_1 = [x_1, x_2, \cdots, x_i]$ 与事件句法库中的句法 $P_2 = [y_1, y_2, \cdots, y_i]$ 进行相似度计算，取最大的相似度为最终相似度，如最终相似度为

100%,则说明该句法已存在于事件句法库中,该文本是一个事件文本。此外,事件文本必须包含触发词 σ,不包含触发词的文本直接判定为非事件文本,综合各方面考虑,本模型最终的相似度计算方法如公式 1 所示:

$$\cos(\theta)=\begin{cases}\dfrac{\sum_{i=1}^{n}(x_i\cdot y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2}\cdot\sqrt{\sum_{i=1}^{n}y_i^2}}, & \sigma\text{ 存在}\\ 0, & \sigma\text{ 不存在}\end{cases}$$

公式(1)

事件句法相似度计算算法如算法 2 所示:

算法 2:事件句法相似度计算

输入:patterns[0..n-1]:包含 n 条事件句法特征(pattern)的数组;sentence:待测文本的语句句法;f1(sentence):输入句法中含有触发词 σ,返回 1,否则返回 0;f2(pattern,sentence):计算两个事件句法的余弦相似度;

输出:cos 事件句法相似度

```
function Cos(patterns[0..n-1]:array of pattern;sentences:text to be tested; f1: function; f2: function): cos;
var
  σ: 触发词识别变量;
  cos:句法相似度;
  temp:临时变量;
begin
  for i←0 to n-1 do
    pattern ← patterns[i]
    σ ← f1 (sentence)
     temp ← f2(pattern,sentence)
     temp ← temp · σ
     if temp > cos then
        cos ← temp
     end if
   return cos
end
```

4 融合句法特征和句法相似度的网络舆情突发事件识别方法

4.1 句法特征和句法相似度在事件识别方面的优势分析

国内外学者的研究已经证明,融入句法分析或句法特征的句法分析统计模型[24]、中文实体关系抽取[25]、答案提取算法[27]、中文复杂名词短语分析[28]等都取得了良好的实验结果,语句的句法特征从语义层面对文本的语法规则和语言形式进行描述[23],句法特征能够表达语句的语义特征。不同于传统的文本特征,句法特征描述了句子中的依存结构、短语结构以及功能,使得事件识别模型在事件识别的过程中更加关注事件语词之间的语义逻辑和依存关系,这对于提升事件识别模型的查全率和查准率有着很大的帮助。事件本身存在一定的语义逻辑和句法结构,使用句法特征表达事件存在着先天的优势,使用句法特征能够使得事件识别的重心从文本内容转移到语义逻辑,从而避开中文文本的种类和数量规模,以到达识别事件的目的。

4.2 网络舆情突发事件识别方法整体设计思路

笔者构建了如图 1 所示的网络舆情突发事件识别模型,模型分为两个部分:①事件句法特征库构造,首先通过网络爬虫获取社交媒体上相关舆情语料训练集文档 $TD=(TD_1,TD_2,\cdots,TD_i)$,然后人工标注出文档 TD_i 的事件集 $E=\{E_1,E_2,\cdots,E_j \mid E_j\in TD_i\}$,接着通过句法特征抽取方法得到事件集 E 对应的句法 $P_m=\{<E_1:P_1>,<E_2:P_2>,\cdots,<E_j:P_j>\}$,按此方法从舆情语料训练集文档中获取所有事件句法,在经过去重、人工修正等操作后,形成舆情语料的事件句法特征库;②待测文本事件识别,首先对待识别文档 $D=(D_1,D_2,\cdots,D_i)$ 进行分句操作,将文档 D_i 切割为由语句所包含的语句集合 $S=\{S_1,S_2,\cdots,S_j \mid S_j\in D_i\}$,接着通过句法特征提取得到文档 D_i 的句法特征集合 $P_n=\{<S_1:P_1>,<S_2:P_2>,\cdots,<S_j:P_j>\}$,$P_n$ 作为待识别文本句法进入到模型中与舆情事件句法库中现有事件句法进行相似度计算,相似度大于或等于模型相似度阈值的待测文本即为事件文本。

4.3 网络舆情突发事件识别方法的实现

4.3.1 网络舆情突发事件的语义标注

对网络舆情突发事件进行语义标注的目的是对舆情文本进行事件标注从而获得一定规模的已知事件,为后续进行事件句法特征库的构造打下基础。笔者在进行事件语义标注时除保留事件三元组包含的主谓宾相关实体,其他实体信息诸如地点实体、时间实体、事件实体等实体信息也同样保留,这样使得获得的事件句法模式更加完整,也提升后续利用句法模式所识别新事件的准确性。

对于事件语义标注,笔者定义了以下几条标注原则:

原则一:所标注的事件文本必须能够从中推导出事件的发生。

原则二:在满足原则一的情况下,所推导出的事件必须是真实发生过或正在发生的事件,例如事件文本中包含否定词语、未来发生词语、可能发生词语、个体

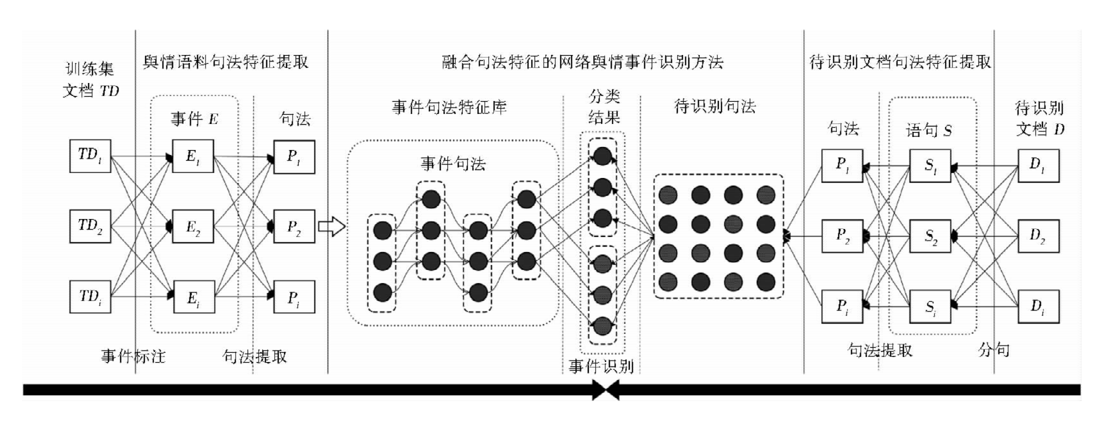


图 1　网络舆情突发事件识别模型

主观推测发生词语时，不算作事件。

原则三：在满足原则一的情况下，事件文本中的时间实体和地点实体属于事件的一部分，应当被标注。

原则四：在满足原则一的情况下，一个事件可作为另一个事件的施事者或者受事者，即事件本身也可以作为一个实体。

4.3.2　基于网络舆情突发事件识别模型的事件识别方法

以4.2所提出的网络舆情突发事件识别模型为核心，对特定领域网络舆情突发事件进行事件识别。事件句法特征库在事件识别方法中承担着事件句法训练集的作用，因此构造完备的事件句法特征库是首要任务；在完成事件句法特征库的构造后，对待测文本进行分句并进行事件识别。

(1)事件句法特征库构造。事件句法特征库构造分为两个子模块：①网络舆情语料采集和语义标注。通过自主编写python爬虫获取相关领域网络舆情突发事件语料，在对数据进行一定的清洗之后，通过4.3.1所提出的网络舆情突发事件语义标注对所采集的语料信息进行事件标注形成舆情事件语料库。②语句词性标注和句法特征提取。利用jieba分词工具对事件语料库中的事件文本依次进行分词和词性标注，之后通过3.1节所提出的面向事件识别的语句句法特征提取方法对事件进行句法特征提取，所获取句法进入到事件句法特征库。为避免产生重复句法特征，需要对新入库的句法特征进行重复判断，若句法重复，则不进行入库处理，并同时采用错误模式纠查反馈机制，利用人工方式减少错误句法的产生。事件句法特征库构造具体流程见图2。其中，为保证分词和词性标注的准确性，除利用jieba分词词典外，还将结合具体的网络舆情突发事件定义该领域的自定义词典，例如在“台风利奇马”事件中，“利奇马”一词本意为越南的一种水果，通过词典定义“台风利奇马”是词性为“名词(n)”的一个词语，使得在分词的过程中，该词语不会被拆分且词性正确。

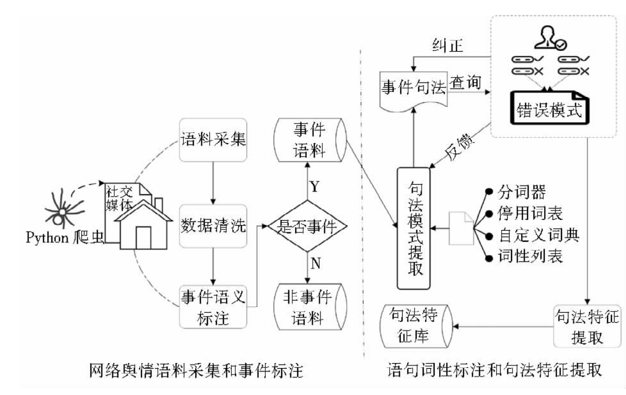


图 2　事件句法特征库构造流程

(2)待测文本分句与事件识别。对待测事件文本进行分句的主要困难在于不清楚待测舆情语料中包含事件文本的位置，由于事先不清楚文本中的事件结构，事件的位置可能存在于半句话或一句话中，事件本身也可以成为另一个事件的一部分元素。对此，笔者采用的方法是对一段待测文本进行重复分句，例如对待测文本“美国再次调查中兴通讯，使其股价应声大跌”可以将其分句为“美国再次调查中兴通讯”“使其股价应声大跌”“美国再次调查中兴通讯，使其股价应声大跌”三段文本。对一段文本重复分句可以识别出其中所包含的所有事件，避免产生遗漏。

文档 D_i 包含大量网络舆情突发事件，首先对其进行分句操作，通过3.1节所提出的面向事件识别的语句句法特征提取方法对待测文本进行句法特征提取，获得文档 D_i 的句法特征集合 $P_n = \{<S_1:P_1>, <S_2:P_2>, \cdots, <S_j:P_j>\}$，$P_n$ 进入到模型中与事件句法特征库中的句法按照算法2进行相似度计算。因网络舆情类别不同，每一个特定类别的网络舆情事件识别模

型都对应着一个特定的相似度阈值 α，最终句法相似度大于或等于相似度阈值 α 即为事件。使得模型的 F1 值最优的句法相似度即为相似度阈值 α，F1 值计算方式如公式 2 所示：

$$F1 = \frac{2PR}{P + R} \qquad \text{公式(2)}$$

5 网络舆情突发事件识别实证分析——以新型冠状病毒肺炎疫情为例

以新型冠状病毒肺炎疫情为例，验证笔者所提出的融合句法特征和句法相似度的网络舆情突发事件识别方法的有效性。

5.1 语料采集与事件标注

通过自主编写 python 爬虫以关键词“新型冠状病毒肺炎疫情”爬取微博平台相关数据，并对数据进行清洗和预处理，形成 3 份不同的网络舆情语料文档 D_1、D_2、D_3。其中，文档 D_1 用以构造事件句法特征库，文档 D_2 用以计算事件识别模型在新型冠状病毒肺炎疫情事例下的相似度阈值 α，文档 D_3 用以检验事件识别模型从未知文本中识别新事件的能力。根据文档功能不同，对文档 D_1、D_2 进行事件语义标注，对 D_3 文档进行重复分句操作。文档 D_1 经过事件语义标注后得到 1 353个事件，再通过句法特征提取，共构造了包含 1 328条有效句法的事件句法特征库，每一条句法代表着一个事件在句法逻辑层面的特征。

文档 D_1 经过事件语义标注后获得的事件以及对应的事件句法特征库如图 3 所示：

中央指导组专家、北京朝阳医院副院长童朝晖接受央视访问时表示，一般两次核酸检测都为阴性，且肺部病变吸收较好的病患才会出院	n-x-n-b-n-v-j-v-n-v-x-a-m-n-v-d-p-n-x-zg-n-v-a-u-n-d-v-n
菏泽第4例新冠肺炎患者治愈出院	n-m-n-v-n
2月9 号上午10点45分，在东明县人民医院接受治疗的东明首例新型冠状病毒感染的肺炎患者出院	m-x-m-t-m-v-x-p-n-v-u-n-b-l-v-u-n
患者李先生1月14日从武汉返回东明老家	n-m-p-n-v-n
患者李先生1月17日出现发热症状	n-m-v-n
患者李先生1月23日在东明县人民医院接受隔离治疗	n-m-p-n-v
患者李先生1月28日被确诊为新型冠状病毒感染的肺炎	n-m-p-v-p-b-l-v-u-n
这么一群人迎难而上，每天背着一个统一的帆布挎包，戴着一只口罩穿梭于社区的大街小巷中	r-m-n-l-x-r-v-m-v-u-n-x-v-u-m-n-v-p-n-u-i-f
2月7日，山东大学齐鲁医院131人的医疗团队奔赴湖北武汉抗疫	m-x-n-m-n-u-n-v-n-v
8日，元宵节深夜，齐鲁医院泌尿外科徐亚楠发布朋友圈消息	m-x-t-x-n-v-n
在这场疫情防控中，除了大家熟知的李文亮医生，还有很多人牺牲在工作岗位上	p-m-n-v-f-x-p-n-v-u-n-x-v-m-n-v-p-n-f
恐怖分子闯进了日内瓦的国际卫生组织总部	n-v-u-n-u-n
恐怖分子受到了保安人员的阻击	n-v-u-n-u-v
武汉中心医院医生李文亮因新冠肺炎病逝	n-p-n-v
李文亮因新冠肺炎病逝	n-p-n-v
环球时报记者6日晚从多个消息源了解到，武汉市中心医院医生李文亮因新冠病毒感染的肺炎于当日病逝	n-m-t-p-m-n-v-x-n-p-n-u-n-p-t-v
2月16日0—24时，31个省（自治区、直辖市）和新疆生产建设兵团报告新增确诊病例2048例	m-x-m-n-x-m-n-x-n-x-n-x-c-n-v-n-m-v
湖北新增病例降至三位数	n-v-n-v-p-n
湖北新增确诊病例349例	n-v-n-m-v
钟南山院士呼吁：解决疫情最快，成本最低的方式就是全中国人民在家隔离两周，这样对全国经济影响最小，对生命健康最有利	n-v-x-v-n-d-x-n-a-u-n-d-a-n-r-v-m-x-r-p-n-v-a-x-p-v-a-d-a
钟南山院士强烈建议全中国人民都在家过春节，不要走亲访友	m-x-m-n-x-n-b-l-v-u-n-v-n
1月23日0-18时，江苏省报告新型冠状病毒感染的肺炎新增确诊病例4例	n-v-n-v-n-b-l-v-u-n-v-n
江苏新增4例输入性新型冠状病毒感染的肺炎确诊病例	n-m-n-p-v-s-p-f-a-n
武汉7名医生在请战书上按下红手印	n-m-b-n-p-a-v
武汉一名新型肺炎患者被成功救治	v-n-u-n-x-d-v-m-x-f-x-d-v-p-m-f-v-n
常驻武汉的患者，已经发热到38℃以上，还执意在一天后回到江苏	x-n-b-l-v-u-n-v-n-x-v-n-x-v-a-l-n-j-v-n
《湖北省新型冠状病毒感染的肺炎诊疗方案（试行第一版）》发布新冠状病毒肺炎中医预防方案	m-x-n-b-l-v-u-n-v-n-x-v-n-x-v
1月23日《湖北省新型冠状病毒感染的肺炎诊疗方案（试行第一版）》发布	m-t-v-u-n-v-j-n-v-n-f-x-n-l-n-v-x-p-n-v-x-m-x-n-v-n-m-v-m-v-f
2月18日下午举行的国务院联防联控机制召开新闻发布会上，国家卫健委新闻发言人米锋介绍，与高点相比，2月17日，单日新增确诊病例首次降至2000例以内	n-v-n-b-v-n-v-n
陕西首开高铁专列运送医疗队援鄂	v-m-n-p-n-v-n
还有一批军队医护人员通过铁路驰援武汉	t-x-n-x-m-u-m-n-d-v-n
今天，空军运-20等8架运输机再飞武汉	m-x-n-v-j-n-v
2月17日，国务院联防联控机制新闻发布会召开	t-x-n-v-j-n
今天，国务院举行联防联控机制发布会	n-l-v-b-l-v-u-n-v-p-x-b-l-n-x-v-x-n-x
新闻发言人介绍新型冠状病毒感染的肺炎统一称谓为“新型冠状病毒肺炎”，简称“新冠肺炎”	t-v-m-x-n-x-n-t-v
现在有很多“小汤山”医院正在建设	t-v-x-n-v-b-l-n-v-j-n-t-v-d-b-l-v-u-n
会上介绍，国务院应对新型冠状病毒肺炎疫情联防联控机制成员现决定将新型冠状病毒感染的肺炎	m-x-n-a-n-p-n-v-u-m-n-v-n-c-n-v-n
2月8日 俄罗斯紧急事务部为中国筹措的第二批医疗防护物资及药品运抵莫斯科	n-v-x-r-q-n-j-v-m-q-x-n-d-m
俄方称，这批物资一共是183立方米，总重量约25吨	t-x-m-x-n-v-m-q-n-p-n-v
今天（2月18日）广东省政府新闻办疫情防控第二十四场新闻发布会在广州举行	n-v-n
钟南山参加广东疫情发布会	n-v-n
广东抗击肺炎疫情	n-l-n-l-v-n-b-x-n-m-v-x-b-l-v-l-u-m-x-r-p-eng-l-v-x-n-u-x

图 3　文档 D_1 所标注事件和对应事件句法特征库

文档 D_2 经过事件语义标注后得到65 个事件和54 个非事件，通过句法特征提取方法获取 119 条句法，这些句法用来确定模型相似度阈值 α。文档 D_2 经过事件语义标注后获得的事件和非事件以及对应的语句句法特征见图 4。

文档 D_3 经过重复分句后形成一个以语句集合表达的待测文档，对语句集合进行相应的句法特征提取形成测试集。文档 D_3 经过分句后形成的语句集合和对应的句法特征见图 5。

5.2 新型冠状病毒肺炎疫情事件识别结果展示与分析

为确定最优相似度阈值 α 的取值，首先通过网络舆情突发事件识别模型依次将文档 D_2 所标注的事件和非事件句法与事件句法特征库中的事件句法进行相似度计算，获得文档 D_2 中所有事件与非事件的句法相似度；接着将相似度阈值 α 按照 0.01 的步长在区间[0,1]中依次取值，直至获得使模型 F1 值最优的 α 取值；最终实验结果如表 1 所示，相似度阈值 α 在[0.89, 1]的范围内就能得到模型的最优结果。通过对比在不同 α 取值下的 P 值、R 值、F1 值走势图（见图 6），可以确定事件识别模型在新型冠状病毒肺炎疫情中的最优相似度值为 0.93，此时对应文档 D_2 实验结果的 F1 值为 0.786、P 值达到 0.713、R 值 0.877。

广西新增新冠肺炎确诊病例8例	n-v-n-v-n-m-v
2020年1月23日0—24时，广西报告新型冠状病毒感染的肺炎新增确诊病例8例	m-x-m-n-x-n-b-l-v-u-n-v-n-m-v
百色确认首例新型冠状病毒病例	n-v-n-b-l-n
科比去世	j-v
2020年过去还不到一个月	m-t-d-v-m
全州女孩大半夜冲浪	n-t-v
我南宁的号都收到了短信	r-n-u-q-d-v-u-n
广西启动一级响应	n-v-m-v
北海爆料	n-v
北海市新增新型冠状病毒感染的肺炎确诊3例病例情况公布	n-v-b-l-v-u-n-v-n-v
其中北海市新增新型冠状病毒感染的肺炎确诊病例3例	r-n-v-b-l-v-u-n-v-n
患者韩某，1月18日到达北海	n-x-m-v-n
患者韩某，19日出现发热、乏力等症状	n-x-m-v-x-a-u-n
我妈妈前两天陪我姥姥去医院	r-n-i-v-r-n-v-n
她姐姐的老板带着她女儿从武汉回南宁了	r-n-u-n-v-u-r-n-p-n-v-n-u
河池爆料	n-v
广西确诊新型肺炎病例	n-v-b-n
河池市首例新型冠状病毒感染的肺炎病例情况公布	n-b-l-v-u-n-v
1月24日凌晨，河池市首例疑似新型冠状病毒感染的肺炎病例确诊	m-t-x-n-v-b-l-v-u-n-v
患者李某，2020年1月13日至1月17日在武汉某医院进行软件开发和维护工作	n-x-m-p-m-p-n-r-n-v-l-c-v
我现在连出门遛狗都得带个口罩	r-t-n-v-d-u-v-q-n
不少武汉逃离到桂林的武汉人住着民宿，逛着景区，过着大年	d-n-v-n-u-n-v-u-n-x-v-u-n-x-u-m
北京市卫健委新闻发言人高小俊介绍，截至2020年2月14日24时，北京市20家定点医院中医药参与救治率为90%	n-l-n-v-x-v-m-n-x-n-m-n-v-m-x
国家中医药管理局医疗救治专家组组长、中国工程院院士、中国中医科学院院长黄璐琦表示，目前湖北地区确诊病例中医药参与率75%以上	n-v-n-v-x-n-x-n-j-n-v-x-t-n-v-n-m-x-f
湖北地区一半以上的确诊病例都使用中医药治疗	n-m-f-u-v-n-d-v-n-v
去年我因为吃感冒药过敏	t-r-c-v-n
去年我得了荨麻疹	t-r-v-n
后来，我去中医院	t-x-r-v-n
今天晚上七政四余高级班讲解健康问题	t-n-m-n-v-a-n
新闻称北京20家定点医院中医药参与救治率90%	n-v-n-m-n-v-m-x
中国疾病预防控制中心新型冠状病毒肺炎应急响应流行病学组	n-v-l-b-l-n-v-n-zg
中国疾病预防控制中心分析7万多名疑似和确诊患者	n-v-l-v-x-m-v-c-v-n
中国疾病预防控制中心发现COVID-19患者中约80.9%为轻中症患者	n-v-l-v-eng-x-m-n-m-x-p-a-n
武汉20位康复医护人员捐血浆	n-m-v-n-v-n
最近中医引起了很多争论	f-j-v-u-m-v

图 4　文档 D_2 所标注事件与非事件和对应语句句法

这是全球最关注的物质，没有之一	r-v-n-d-v-u-n-x-v-r
没有之一	v-r
大水里才有大鱼	n-f-d-v-n
以下五点你必须认真看完	f-m-r-d-a-v
1.进度。	m-d-x
新冠肺炎	n
临床一共三期	v-j-t
图2：参与YM研发的公司进度汇总，A股受益公司	n-m-x-v-eng-j-u-n-d-n-x-n-v-n
图2	n-m
参与YM研发的公司进度汇总	v-eng-j-u-n-d-n
A股受益公司	n-v-n
4月9号查到的信息——2期临床试验已处于预注册状态，也就是说特事特办以及紧急状态时使用	m-v-u-n-x-m-n-d-v-n-x-l-n-c-l-n-v
4月9号查到的信息——2期临床试验已处于预注册状态	m-v-u-n-x-m-n-d-v-n
也就是说特事特办以及紧急状态时使用	l-n-c-l-n-v
国际劳工组织：新冠疫情已影响全球超八成劳动人口	n-x-n-d-v-n-v-m-v-n
国际劳工组织	n
新冠疫情已影响全球超八成劳动人口	n-d-v-n-v-m-v-n
国际劳工组织7日发布的报告显示，在全球33亿劳动人口中，已有81%受到#新冠肺炎# 疫情影响，其工作场所被全部或部分关闭	n-m-v-u-n-v-x-p-n-m-v-n-f-x-v-m-x-v-x-n-x-n-v-x-r-v-n-p-n-c-n-v
国际劳工组织7日发布的报告显示	n-m-v-u-n-v
在全球33亿劳动人口中	p-n-m-v-n-f
已有81%受到#新冠肺炎# 疫情影响	v-m-x-v-x-n-x-n-v
其工作场所被全部或部分关闭	r-v-n-p-n-c-n-v
在全球33亿劳动人口中，已有81%受到#新冠肺炎# 疫情影响，其工作场所被全部或部分关闭	p-n-m-v-n-f-x-v-m-x-v-x-n-x-n-v-x-r-v-n-p-n-c-n-v
已有81%受到#新冠肺炎# 疫情影响，其工作场所被全部或部分关闭	v-m-x-v-x-n-x-n-v-x-r-v-n-p-n-c-n-v
报告预测，疫情将使今年第二季度全球劳动人口总工时缩减6.7%，相当于1.95亿名全职雇员失业	n-v-x-n-d-v-t-m-n-v-n-v-m-x-v-m-n
疫情将使今年第二季度全球劳动人口总工时缩减6.7%，相当于1.95亿名全职雇员失业	n-d-v-t-m-n-v-n-v-m-x-v-m-n
报告预测	n-v
疫情将使今年第二季度全球劳动人口总工时缩减6.7%	n-d-v-t-m-n-v-n-v-m-x
相当于1.95亿名全职雇员失业	v-m-n
新华网	n
工人日报的微博投票	n-u-a-n
早，#新冠肺炎#疫情下，易北爱乐乐团演奏的布拉姆斯第一，在这个艰难的时刻向大家传递勇气和力量	a-x-n-x-n-f-x-n-v-u-n-m-x-p-r-a-u-n-p-n-v-n-c-n
#新冠肺炎#疫情下，易北爱乐乐团演奏的布拉姆斯第一，在这个艰难的时刻向大家传递勇气和力量	x-n-x-n-f-x-n-v-u-n-m-x-p-r-a-u-n-p-n-v-n-c-n
易北爱乐乐团演奏的布拉姆斯第一，在这个艰难的时刻向大家传递勇气和力量	n-v-u-n-m-x-p-r-a-u-n-p-n-v-n-c-n
早	a
#新冠肺炎#疫情下	x-n-x-n-f
易北爱乐乐团演奏的布拉姆斯第一	n-v-u-n-m
在这个艰难的时刻向大家传递勇气和力量	p-r-a-u-n-p-n-v-n-c-n
再大风雨只要有家都可以安然度过，风雨之后一定会再现彩虹	d-i-c-r-d-c-n-v-x-n-f-d-v-n
再大风雨只要有家都可以安然度过	d-i-c-r-d-c-n-v

图 5　文档 D_3 分句结果和对应语句句法

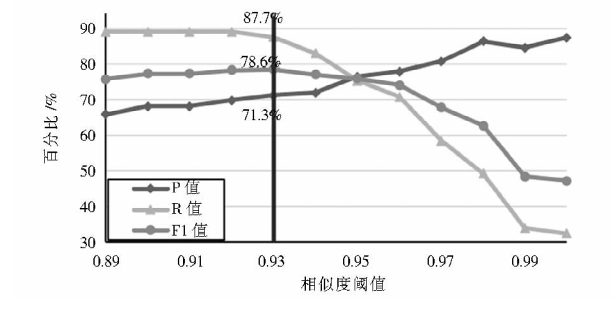


图 6　P 值、R 值、F1 值走势

通过文档 D_2 确定了模型在新型冠状病毒肺炎疫情下的最优相似度阈值 α 为 0.93，接下来模型以相似度阈值为 0.93 的情况对文档 D_3 进行事件识别，最终从文档 D_3 识别出 160 个事件、30 个非事件，部分事件识别结果见表 2。事件识别模型在文档 D_3 的 F1 值达到了 0.848，P 值达到了 0.769，R 值达到了 0.946，结果见表 3。

表 1 不同阈值对应的 P 值、R 值、F1 值

相似度阈值	P 值	R 值	F1 值
1	0.875	0.323	0.472
0.99	0.846	0.338	0.484
0.98	0.865	0.492	0.627
0.97	0.809	0.585	0.679
0.96	0.780	0.708	0.742
0.95	0.766	0.754	0.760
0.94	0.720	0.831	0.771
0.93	0.713	0.877	0.786
0.92	0.699	0.892	0.784
0.91	0.682	0.892	0.773
0.90	0.682	0.892	0.773
0.89	0.660	0.892	0.758

表 2 从文档 D_3 中识别事件结果(部分)

编号	识别事件
1	4 月 10 日 0 时至 24 时,广州市报告新增确诊病例 4 例,其中境外输入病例 1 例(入境口岸排查发现),境外输入关联病例 3 例(2 例为密切接触者排查发现、1 例为医疗机构发热门诊排查发现)
2	参与 YM 研发的公司进度汇总
3	讲述了国家/个人应对危机的 12 个步骤
4	4 月 9 号查到的信息——2 期临床试验已处于预注册状态,也就是说特事特办以及紧急状态时使用
5	据韩国中央防疫对策本部 12 日消息,截至当天 0 时,国内共有 111 例#新冠# 肺炎治愈后复发的病例
6	国际劳工组织:新冠疫情已影响全球超八成劳动人口
7	国际劳工组织 7 日发布的报告显示,在全球 33 亿劳动人口中,已有 81% 受到#新冠肺炎# 疫情影响,其工作场所被全部或部分关闭
8	国内共有 111 例#新冠# 肺炎治愈后复发的病例
9	广州一美食店隐瞒客人堂食多人确诊
10	在全球 33 亿劳动人口中,已有 81% 受到#新冠肺炎# 疫情影响,其工作场所被全部或部分关闭
11	已有 81% 受到#新冠肺炎# 疫情影响,其工作场所被全部或部分关闭
12	报告预测,疫情将使今年第二季度全球劳动人口总工时缩减 6.7%,相当于 1.95 亿名全职雇员失业
13	剑桥大学研究显示,意大利疫情可能与德国、新加坡有关,新冠根源病毒在来自美国、澳洲的病例中大量出现,在武汉并不常见
14	工业化国家平均税率是 22.5%,川普大手笔减税将大幅降低其国内企业经营成本,刺激信贷市场的繁荣,吸引制造业和资本回流美国
15	然而#新冠肺炎# 黑天鹅考验各国基础建设与管理能力,如今美元贬值压力远高于往昔,证明人算不如天算
16	吉林省公布新增 1 例境外输入确诊病例
17	为防控疫情 泰国曼谷宣布暂时禁酒
18	欧盟成员国财长会议 4 月 9 日达成协议,将实施总额为 5400 亿欧元的大规模救助计划
19	吉林一输入病例四次核酸检测均为阴性
20	该确诊病例为吉林市人,3 月 16 日从泰国曼谷飞抵广州

表 3 事件识别模型在文档 D_3 中实验结果

文档	P 值	R 值	F1 值
D_3	0.769	0.946	0.848

5.3 方法测评

笔者所提出的网络舆情突发事件识别模型主要创新点在于以句法特征来代替文本特征,以此来解决事件识别所面临的网络舆情突发事件初期语料文本短缺、训练集不足的问题,有效降低了事件识别的维度,并且在模型中以"相邻相同词性合并"的方式再次减少了句法冗余情况。为验证在以上两个方面所做出创新的有效性,笔者选用相邻相同词性不合并的网络舆情突发事件识别模型和基于文本相似度的事件识别方法作为对照进行比较。基于文本相似度的事件识别方法通过文本分词后构造中文文本向量进行相似度计算,该对照组除不使用句法特征表示事件特征外,其余计算步骤与笔者所提出的网络舆情突发事件识别模型步骤保持一致,3 种方法均使用文档 D_2 作为测试集进行试验。

在使用相同的训练集和测试集的情况下,3 种事

件识别方法实验结果如表 4 所示,其中笔者所提出的网络舆情突发事件识别模型表现最优,F1 能够达到 0.786,证明模型使用句法表示事件特征的合理性以及采用"相同相邻词性合并"的有效性。根据实验结果绘制出在相似度阈值 α 不同取值情况下 3 种不同事件识别方法结果的比较,结果如图 7 所示。通过图 7 可以看出在这 3 种方法中,基于文本相似度的事件识别方法实验结果最差,最优 F1 值也只有 0.657,造成这种较差实验结果的主要原因是因为训练集规模不大,所使用的训练集一共只包含 1 353 个事件,相较于传统的中文文本训练集,笔者所使用的训练集规模较小,但这也从另一个方面证明了在训练集规模较小的情况下,笔者所提出的网络舆情突发事件识别模型的优越性;同时,网络舆情突发事件识别模型在同等情况下实验结果优于相邻相同词性不合并的网络舆情突发事件识别模型,造成这种现象的主要原因在于网络舆情突发事件识别模型合并相同词性后,降低了句法向量的维度,减少了因词性冗余造成的不必要计算,从而提升了模型 F1 值。

表 4　3 种不同事件识别方法的实验结果

事件识别方法	最优相似度	P	R	F1
网络舆情突发事件识别模型	0.93	0.713	0.877	0.786
相邻相同词性不合并的网络舆情突发事件识别模型	0.86	0.617	0.892	0.730
基于文本相似度的识别方法	0.63	0.613	0.708	0.657

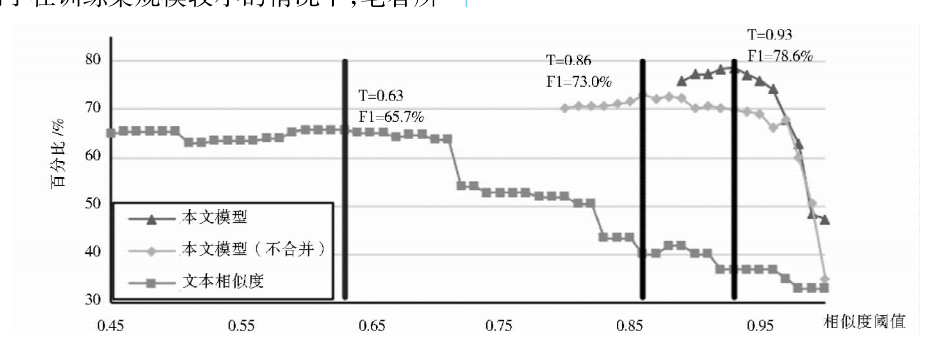


图 7　3 种不同事件识别方法的比较

6　结论与讨论

面对网络舆情突发事件,从社交媒体中快速准确地识别出能够反映网络舆情态度的事件,对政府舆情管理和相关部门决策部署有着很重要的意义。笔者从文本句法特征视角出发,认为事件句法特征能够代替文本特征表示事件,以此作为识别网络舆情突发事件的一个突破口,并在此基础上提出了融合句法特征和句法相似度的网络舆情突发事件识别方法。相较于文本特征,句法特征能够有效的降低中文文本维度,将由数以万计的汉字所构成的语句降维为由数十个词性所构成的句法,大大降低了向量维度,并在此基础上将句法中相邻相同词性合并,再次降低了句法的种类。因此,即使在训练集语料规模较小的情况下,模型仍能表现出较好的事件识别结果。

在以新型冠状病毒肺炎疫情为例的网络舆情突发事件中,笔者所提出的网络舆情突发事件识别模型在最优相似度阈值为 0.93 的情况下,从一段待测文本中识别出事件和非事件,F1 值达到了 0.848。在使用相同的训练集和测试集的情况下,笔者所提出的方法优于相邻相同词性不合并的网络舆情突发事件识别模型和基于文本相似度的事件识别方法,证明了在训练集语料规模较小的情况下,使用句法相似度进行事件识别要优于使用文本相似度,同时也证明了笔者采用相同相邻词性合并策略的合理性,为网络舆情突发事件识别提供了一种新的思路。

从社交媒体文本中识别网络舆情突发事件,对于网络舆情特征分析和演化分析有着十分重要的意义。在后期的研究中,笔者将利用本研究所提出的网络舆情突发事件识别方法识别网络舆情所包含的未知事件,并基于所识别的事件对网络舆情进行进一步的分析和研究。

参考文献:

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[R]. 北京:中国互联网络信息中心, 2019.
[2] DING X, LI Z, LIU T, et al. ELG: an event logic graph[J/OL]. arXiv preprint arXiv: 1907.08015[2021－04－11]. https://arxiv.org/abs/1907.08015.
[3] AGUILAR J, BELLER C, MCNAMEE P, et al. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards[C]//Proceedings of the second workshop on events: definition, detection, coreference, and representation. USA: Association for Computational Linguistics, 2014: 45－53.
[4] 吴刚. 基于主题的中文事件抽取技术研究及应用[D]. 苏州: 苏州大学, 2009.
[5] 项威,王邦. 中文事件抽取研究综述[J]. 计算机技术与发展,

2020(1): 1－9.
[6] CHUNG S, TIMBERLAKE A. Tense, aspect and mood[M]// Language typology and syntactic description. Cambridge: Cambridge University Press, 1985:202－258.
[7] DODDINGTON G R, MITCHELL A, PRZYBOCKI M A, et al. The automatic content extraction(ACE) program tasks, data, and evaluation[C]//Proceedings of the international conference on language resources and evaluation. Portugal: European Language Resources Association, 2004:837－840.
[8] 高强,游宏梁. 事件抽取技术研究综述[J]. 情报理论与实践,2013,36(4):114－117,128.
[9] 李章超,李忠凯,何琳.《左传》战争事件抽取技术研究[J]. 图书情报工作,2020,64(7):20－29.
[10] 贺瑞芳,段绍杨. 基于多任务学习的中文事件抽取联合模型[J]. 软件学报, 2019, 30(4): 1015－1030.
[11] 俞琰. 基于隐马尔可夫模型的招聘网络信息抽取[J]. 北京电子科技学院学报,2008,16(4):93－98.
[12] 李响,杨小琳,魏勇,等. 基于支持向量机的新闻事件类型识别[J]. 地理信息世界,2019,26(2): 73－78.
[13] 刘忠宝,党建飞,张志剑.《史记》历史事件自动抽取与事理图谱构建研究[J]. 图书情报工作,2020,64(11):116－124.
[14] 尉永清,杨玉珍,费绍栋,等. 融合用户情感的在线突发事件识别研究[J]. 情报理论与实践,2015,38(2):92－96.
[15] 武澎,王恒山,刘奇,等. 微博中突发事件信息发布者被“加关注”的阈值模型研究[J]. 情报杂志,2012,31(11):11－13,34.
[16] 刘雅姝,张海涛,徐海玲,等. 多维特征融合的网络舆情突发事件演化话题图谱研究[J]. 情报学报,2019,38(8):798－806.
[17] 兰月新. 突发事件网络衍生舆情监测模型研究[J]. 现代图书情报技术,2013(3):51－57.
[18] 兰月新,曾润喜. 突发事件网络舆情传播规律与预警阶段研究[J]. 情报杂志,2013,32(5):16－19.
[19] 张玉亮. 基于发生周期的突发事件网络舆情风险评价指标体系[J]. 情报科学,2012,30(7):1034－1037,1043.
[20] 陈思菁,李纲,毛进,等. 突发事件信息传播网络中的关键节点动态识别研究[J]. 情报学报,2019,38(2):178－190.
[21] 李纲,徐伟,王馨平. 基于事件要素的组合模型微博热点事件摘要提取[J]. 图书情报工作,2018,62(1):96－105.
[22] 夏立新,陈健瑶,余华娟. 基于事理图谱的多维特征网络舆情事件可视化摘要生成研究[J]. 情报理论与实践,2020,43(10):157－164.
[23] 张宁,朱礼军. 中文问答系统问句分析研究综述[J]. 情报工程,2016,2(1):32－42.
[24] 袁里驰. 基于依存关系的句法分析统计模型[J]. 中南大学学报(自然科学版),2009,40(6):1630－1635.
[25] 郭喜跃,何婷婷,胡小华,等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报,2014,28(6):183－189.
[26] 徐飞,叶文豪,宋英华. 基于 BiLSTM-CRF 模型的食品安全事件词性自动标注研究[J]. 情报学报,2018,37(12):1204－1211.
[27] 胡宝顺,王大玲,于戈,等. 基于句法结构特征分析及分类技术的答案提取算法[J]. 计算机学报,2008(4):662－676.
[28] 陈永波,汤昂昂,姬东鸿. 中文复杂名词短语依存句法分析[J]. 计算机应用研究,2015,32(6):1617－1620.

作者贡献说明:

陈健瑶:提出论文思路、撰写论文、完成实验;
翟姗姗:论文研究框架修改;
夏立新:论文修改与指导;
刘德印:数据爬取和数据整理。

Research on Network Public Opinion Emergency Recognition Method Based on Syntactic Features and Syntactic Similarity

Chen Jianyao Zhai Shanshan Xia Lixin Liu Deyin
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] This study aims to identify events from the text of sudden network public opinion quickly and accurately. [**Method/process**] This paper proposed a method to identify network public opinion emergencies by integrating syntactic features and syntactic similarity. An event oriented syntactic feature extraction method was proposed based on syntactic features. Event syntactic feature database was constructed by using event semantic annotation and syntactic feature extraction methods. The network public opinion emergencies were identified by calculating the syntactic similarity between the text to be tested and the syntax database. [**Result/conclusion**] Taking the novel coronavirus pneumonia epidemic as an example, the optimal similarity of the network public opinion emergency identification method proposed by the author is 0.93 in this public opinion. 160 events and 30 non events are identified from a new text under this similarity, and the F1 value reaches 0.848. Through the method evaluation, it is proved that the proposed method is effective in using syntactic similarity to identify events and merge the same adjacent parts of speech.

Keywords: internet public opinion event identification syntax features syntactic similarity